

The InChI Project

Stephen Heller
InChI-Trust Project Director
steve@inchi-trust.org

The main web sites for the IUPAC InChI project are:

<http://www.iupac.org/inchi>

and

<http://www.inchi-trust.org>

2/5/2013

Slides are available at <http://www.hellers.com/steve/pub-talks/nih-fda-usp-2-5-13/frame.htm>

InChITRUST

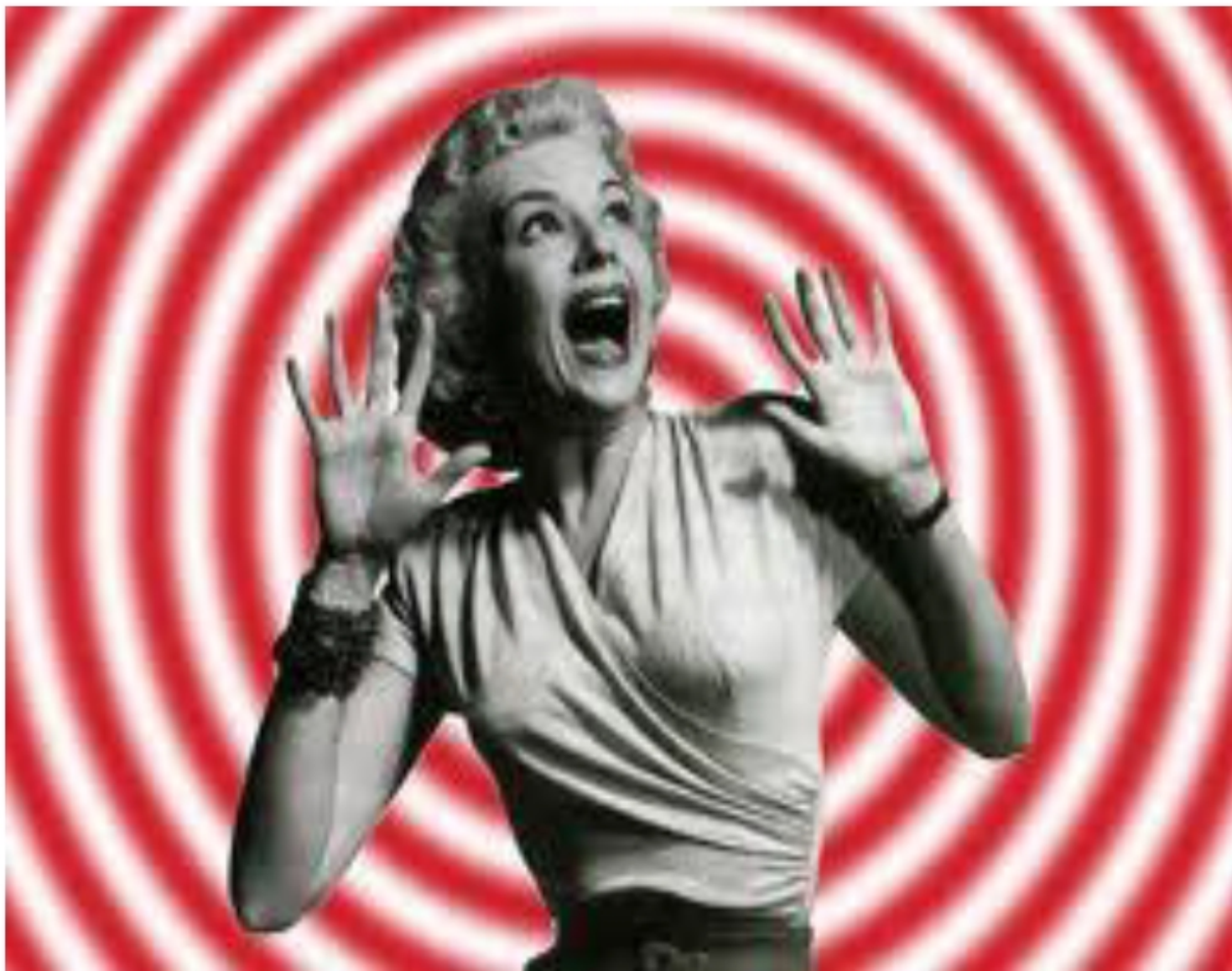


Disclaimer

These slides were made from 100% recycled electrons

InChITRUST





“No, no, not
another
structure
standard!!!”

Why InChI? - Too Many Good and Excellent Identifiers (“Standards”)

Structure diagrams

- various conventions
- contain ‘too much’ information

Connection Tables/Notations

- MolFiles, SDF, SMILES, ROSDAL, ...

Pronounceable names (and mostly unpronounceable) and mostly complex names

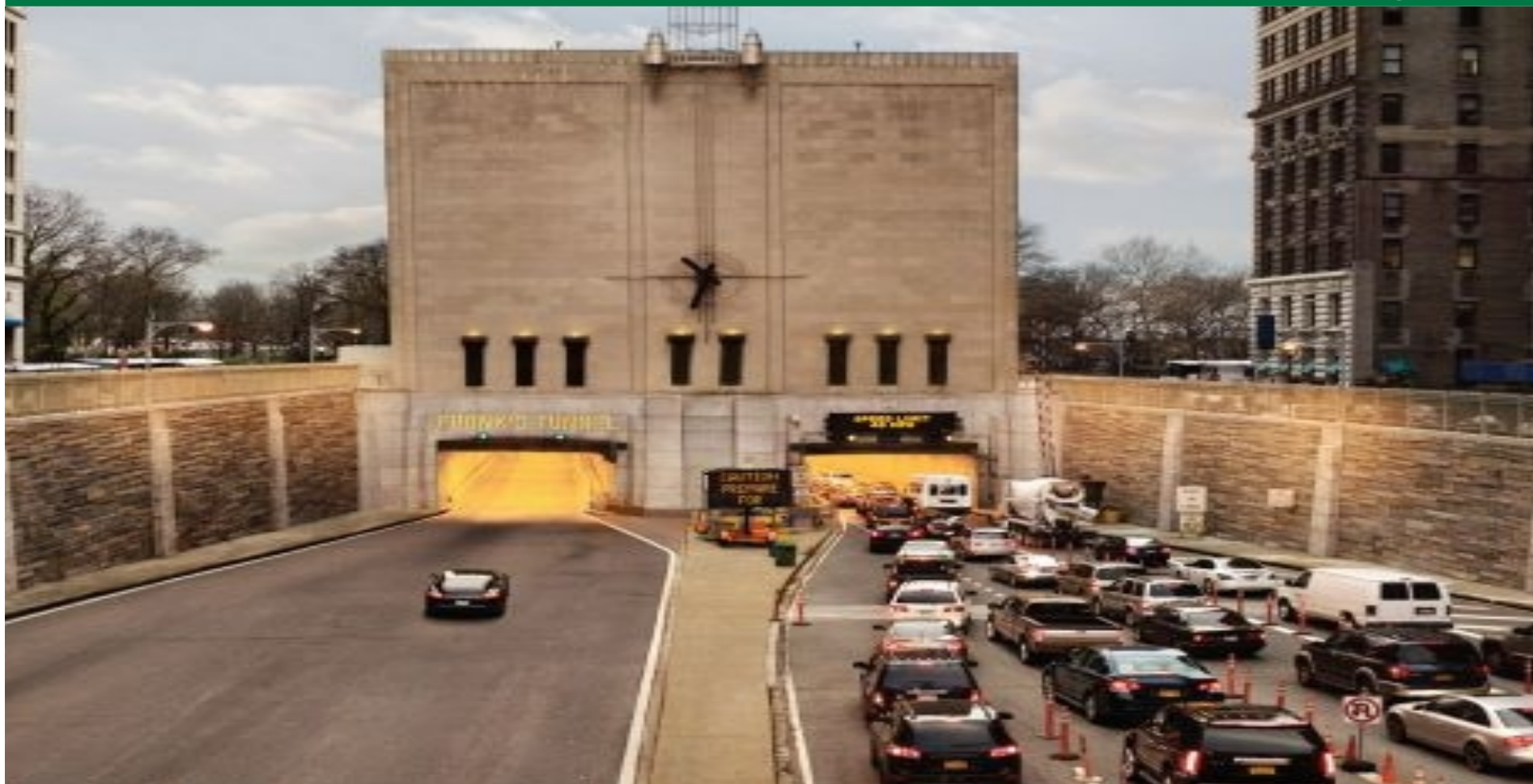
- IUPAC, CAS 8th CI name, CAS 9th CI name, trivial, trade, WHO INN

(Dumb) Index Numbers

- EINECS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN,

EC, ChemSpider ID, REACH, MFCD#, ...

Too many “standards” actually slow things down and make getting to the information you want and need take a lot longer than it can be with InChI, as one can see from the next slide...



InChI

All other standards

What is InChI ?

The IUPAC International Chemical Identifier, or InChI, is a non-proprietary, machine-readable string of symbols which enables a computer to represent the compound in a completely unequivocal manner.

InChIs are produced by computer from structures drawn on-screen with existing structure drawing software, and the original structure can be regenerated from an InChI with appropriate software.

What “*is*” the InChI standard?

The InChI standard programmed into the **algorithm** is a arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., methane). In cases where there is not, one representation is chosen. As long as the arbitrarily chosen representation is properly programmed, one will always get the **SAME** result using it – which is what a standard is!

InChI characteristics

Consensus

Technical competence

Political and technical cooperation

Precompetitive collaboration

No competition with commercial products

No mission creep

IUPAC blessing & rapid IUPAC acceptance

Excellent understanding of what the Internet and how it can be effectively used in Chemical Information

Vision of the future

InChITRUST



While InChI is an Open Source, public domain, system for creating a unique computer-readable identifier (“name”) it is NOT a registry system. InChI’s are created only by those who choose to adopt and use the **algorithm. Registry systems which index the literature are complementary to any InChI databases that anyone creates.**

InChI is not a replacement for any existing internal structure representations. InChI is **IN ADDITION to what one uses internally. Its main value to most organizations is in **LINKING** information**

Critical words/phrases for InChI

Link

Addition; not replacement

Algorithm

No bureaucracy

**PS. Remember the above;
the rest of the lecture is just commentary**

InChITRUST

The logo for InChI TRUST is centered at the bottom of the slide. It features the text "InChI" in a bold, white, sans-serif font, followed by "TRUST" in a regular weight of the same font. Behind the text is a faint, light green hexagonal emblem containing a stylized chemical structure, including what appears to be a balance scale and molecular components.

InChI is for computers

An InChI string is not directly intelligible to the normal human reader. Like Bar Codes, InChIs are not designed to be read by humans.

Or, put another way – never send a human to do a machine's job!

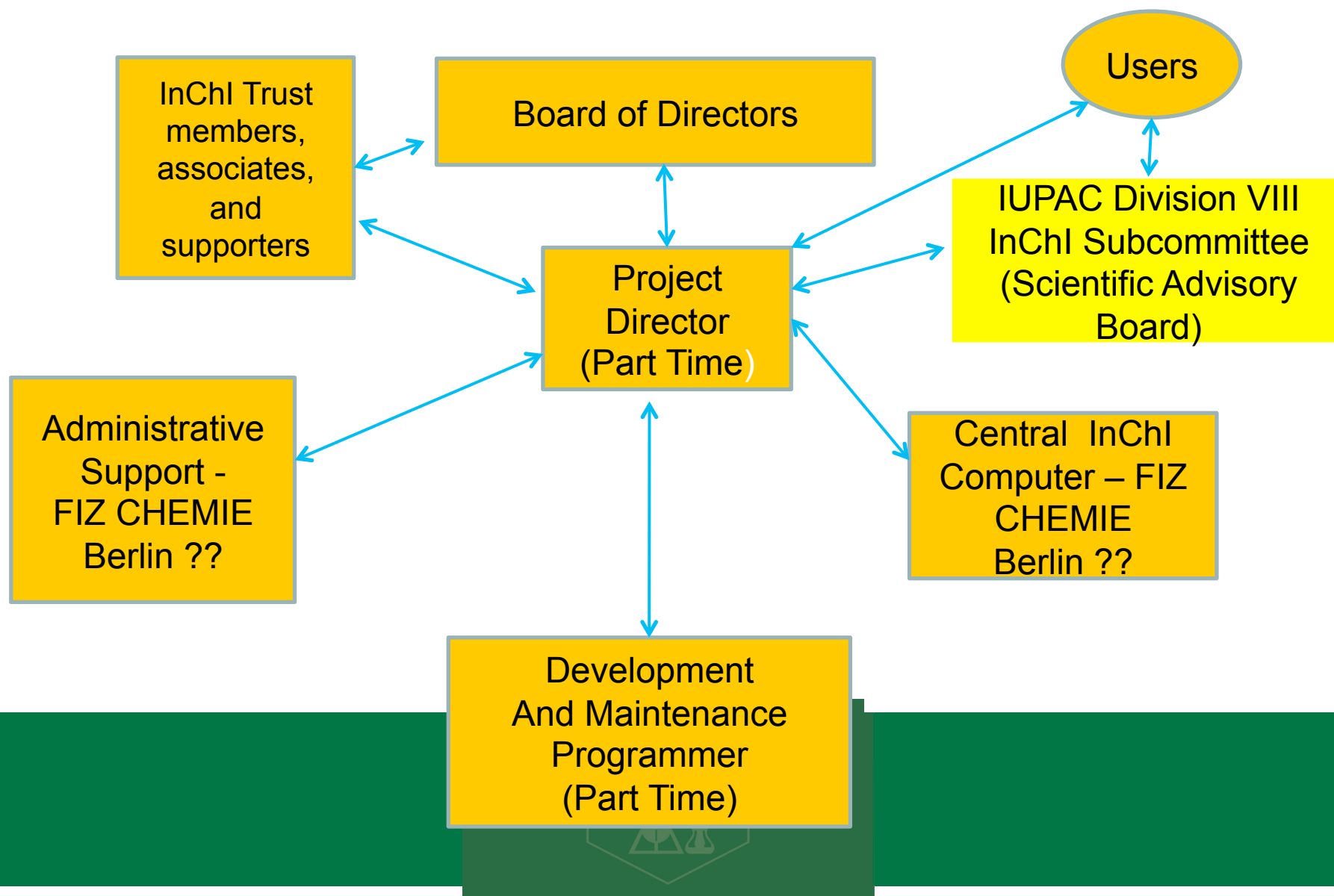
The InChI Trust

To function and succeed, InChI had to become personality independent. InChI had to be “institutionalized”. If the work of this project was to be enduring it needed to be turned over to an entity that would ensure its ongoing activities and be acceptable to the community. It was concluded that a not-for-profit organization would best fit the ongoing and future project needs. Thus the decision to create and incorporate the "InChI Trust" as a UK charity.

InChITRUST

The logo features a green hexagon containing a white chemical structure, which is a representation of a chemical balance scale.

InChI Trust Organization



**Total number of Members,
Associate Members, and (non
paying) Supporters - >50**

InChI Trust Freeloaders

Unfortunately, too numerous to list.

Too many people and organizations don't seem to believe that being part of the chemical information community means supporting the community and paying their "fair share" for the many things they get.

InChI Staff and Collaborators

The InChI project has had the unusual perfect “good storm” of cooperation and support. It is a truly international project with programming in Moscow, computers in Germany, incorporated in the UK, and a project director in the USA. Collaborators from over a dozen countries, from academia, Pharma, publishers, and the chemical information industry, have all offered senior scientific staff to develop the InChI standard.

Why InChI is a success

1. Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be **LINKED** to and combined with other content on the Internet. InChI provides an excellent ROI (return on investment)
2. InChI is a public domain **algorithm** that anyone, anywhere can freely use. And they sure use it!

Success is uncoerced adoption

InChITRUST

The logo for InChI Trust, featuring a stylized chemical structure within a hexagon, with the text 'InChI' in a bold, sans-serif font and 'TRUST' in a smaller, all-caps, sans-serif font.

Bypassing IUPAC procedures

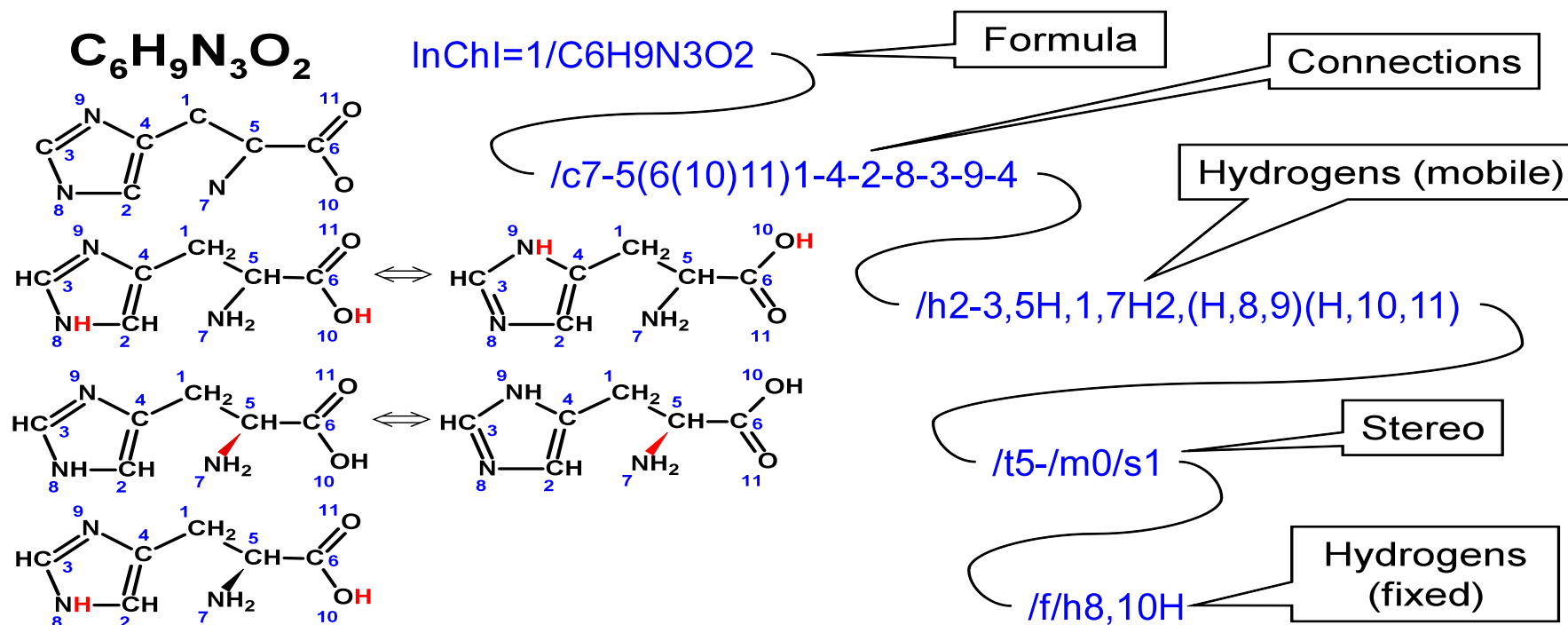
The usual very, lengthy IUPAC approval process was hijacked and sped up by sending the IUPAC bureaucracy, not a white paper with InChI rules, but rather unreadable and unintelligible C code.

How did InChI succeed?

This project was the perfect “good” storm. The project came about in 1999 when Steve Heller retired and his wife threatened him with divorce unless he found some to do. (Yes, behind every successful project is a woman.) IUPAC discovered that nomenclature was for 20th, not 21st century. NIST, the US standards agency, needed a way to represent and link the structures from its standard property databases. The Internet (web 2.0) was taking off enabling silos and islands of information to be linked and searched if only there was a linking element.

Publishers and database producers realized their information would be more valuable (i.e., they could sell more to more people) if only there was a way to link chemical structures from all the diverse resources on the Internet. With no funds to support the project, IUPAC needed the private sector to pay for the short and long term project needs. Lastly, the decentralized structure and hands-off management of the project enabled all the expert egos to be satisfied by putting everyone in charge of what they do best and giving them the final say - allowing for proper, scientific, bottom-up decisions.

InChI Layers: L-Histidine



InChI=1/C6H9N3O2/c7-5(6(10)11)1-4-2-8-3-9-4/h2-3,5H,1,7H2,(H,8,9)(H,10,11)/t5-/m0/s1/f/h8,10H

InChIKey=HNDVDQJCIGZPNO-QLMCEAFFNA-N **InChIKey=HNDVDQJCIGZPNO-YFKPBYRVSA-N**

L-Histidine InChIKey search

Standard InChIKey:

HNDVDQJCIGZPNO-YFKPBYRVSA-N

3,010 hits on Google

InChITRUST

The logo for InChI TRUST, featuring a green hexagon with a white chemical structure inside, and the text "InChI" in white and "TRUST" in green.

Cholesterol InChI search

(using notations from CHEBI website, ID=16113)

InChI:

1S/C27H46O/c1-18(2)7-6-8-19(3)23-11-12-24-22-10-9-20-17-21(28)13-15-26(20,4)25(22)14-16-27
(23,24)5/h9,18-19,21-25,28H,6-8,10-17H2,1-5H3/t19-,21+,22+,23-,24+,25+,26+,27-/m1/s1

818,000 hits on Google

Standard InChIKey:

HVYWMOMLDIMFJA-DPAQBDIFSA-N

56,100 hits on Google

**InChI**TRUST



Henry Rzepa's QR (quick response) smartphone app for InChI

InChIKey: VAYOSLLFUXYJDT-RDTXWAMCSA-N

InChI=1/C20H25N3O/c1-4-23(5-2)20(24)14-9-16-15-7-6-8-17-19(15)13(11-21-17)10-18(16)22(3)12-14/h6-9,11,14,18,21H,4-5,10,12H2,1-3H3/t14-,18-/m1/s1

Lysergic Acid Diethylamide



Current IUPAC Working Groups & Projects

In Progress:

Organometallics
InChI Resolver
Electronic/Excited States
New API

Completed:

Revised FAQ's from Cambridge- Nick Day/Peter Murray-Rust
InChI Certification Suite
Version 1.04 released – 9/11
Markush (contract to be signed when funded)
Polymers/Mixtures
RInChI – InChI for Reactions

Started/To be started in 2013:

InChI teaching/educational materials
Large Molecules
Material Science
Inorganics
Redesign of Handling of Tautomerism



InChITRUST

The Future

InChI has become mainstream for publishers, databases providers, and software developers. Over the next 5-10 years, publishers will use data mining to create both better abstracts, useful indexing, and concept terms. Search engines will be able to search for appropriate text and structures and direct users to the original (fee or free/Open Access/Open Data) sources.

Summary

**If you are not part of the
solution; you are part of the
precipitate**

Acknowledgements

(Primarily members for the IUPAC InChI subcommittee and associated InChI working groups)

Steve Bachrach, Colin Batchelor, John Barnard ,Evan Bolton, Steve Boyer, Steve Bryant, Szabolcs Csepregi ,Rene Deplanque, Gary Mallard, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Bill Milne, Gerry Moss, Peter Murray-Rust, Heike Nau , Marc Nicklaus, Carmen Nitsche, Matthias Nolte , Igor Pletnev, Josep Prous, Peter Murray-Rust, Hinnerk Rey, Ulrich Roessler, Roger Schenck , Martin Schmidt, Steve Stein, Peter Shepherd, Markus Sitzmann ,Chris Steinbeck, Keith Taylor, Dmitrii Tchekhovskoi, Bill Town, Wendy Warr, Jason Wilde, Tony Williams, Andrey Yerin.

Special Acknowledgement: Ted Becker& Alan McNaught for their vision and leadership of the future of IUPAC nomenclature.